基因组学数据分析 第三次作业

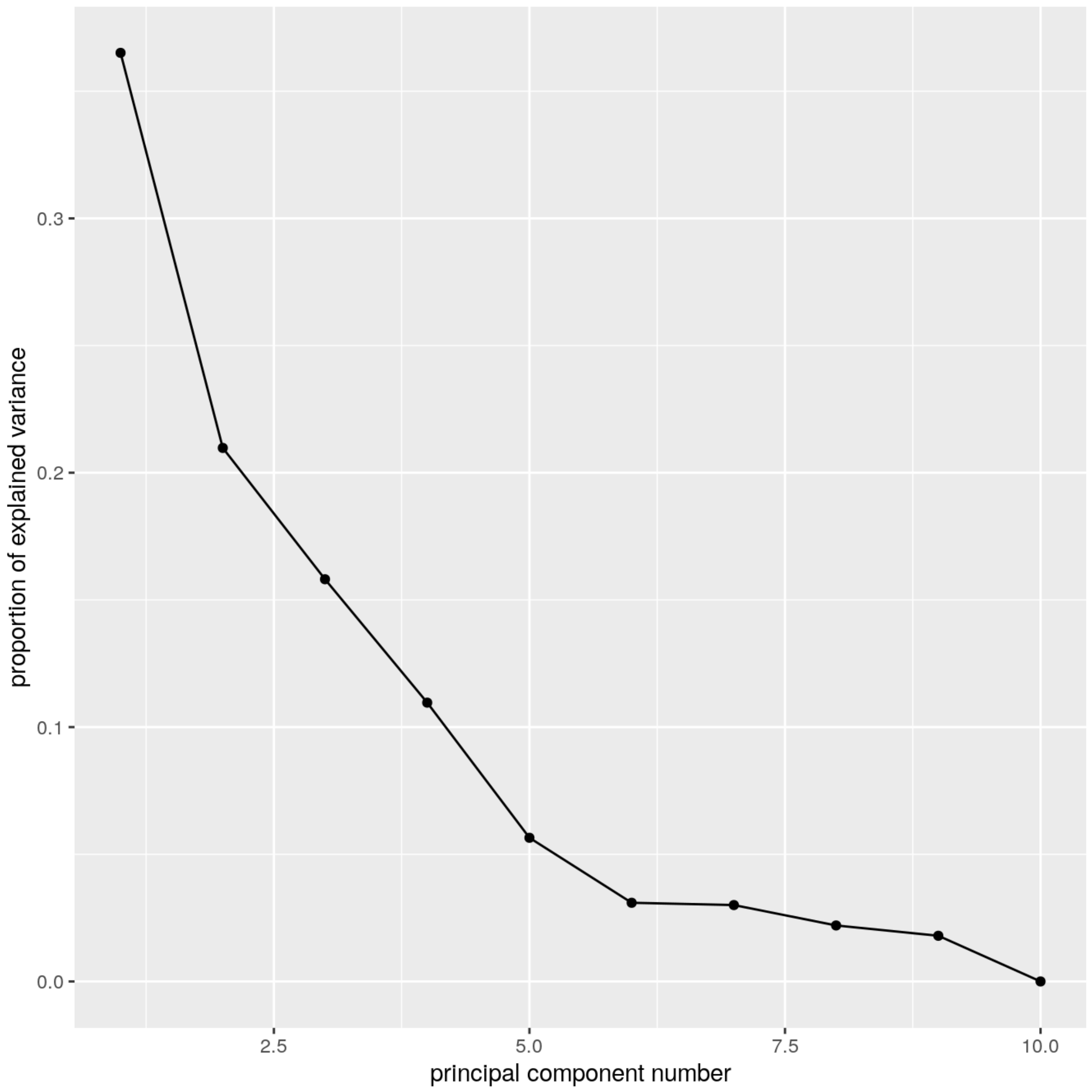
主成分分析法（PCA）分析流程

数据读入与整理

```r
library(ggplot2)
library(dplyr)
genes_expression <- read.delim(
    "./GSE106688_genes.fpkm_table.txt",
    header = TRUE
)
rownames(genes_expression) <- genes_expression[, 1]
colnames(genes_expression) <- c(
    "gene_id", "hESC R1", "hESC R2", "MES R1", "MES R2",
    "CP R1", "CP R2", "CM R1", "CM R2", "Fetal R1", "Fetal R2"
)
genes_expression <- genes_expression[, -1] # 首列作为列名
genes_expression <- genes_expression[
    rowSums(genes_expression) > 0.01,
] # 筛去全为0的行
genes_expression <- genes_expression[
    apply(genes_expression, 1, var) != 0,
] # 筛去方差为0的行
```

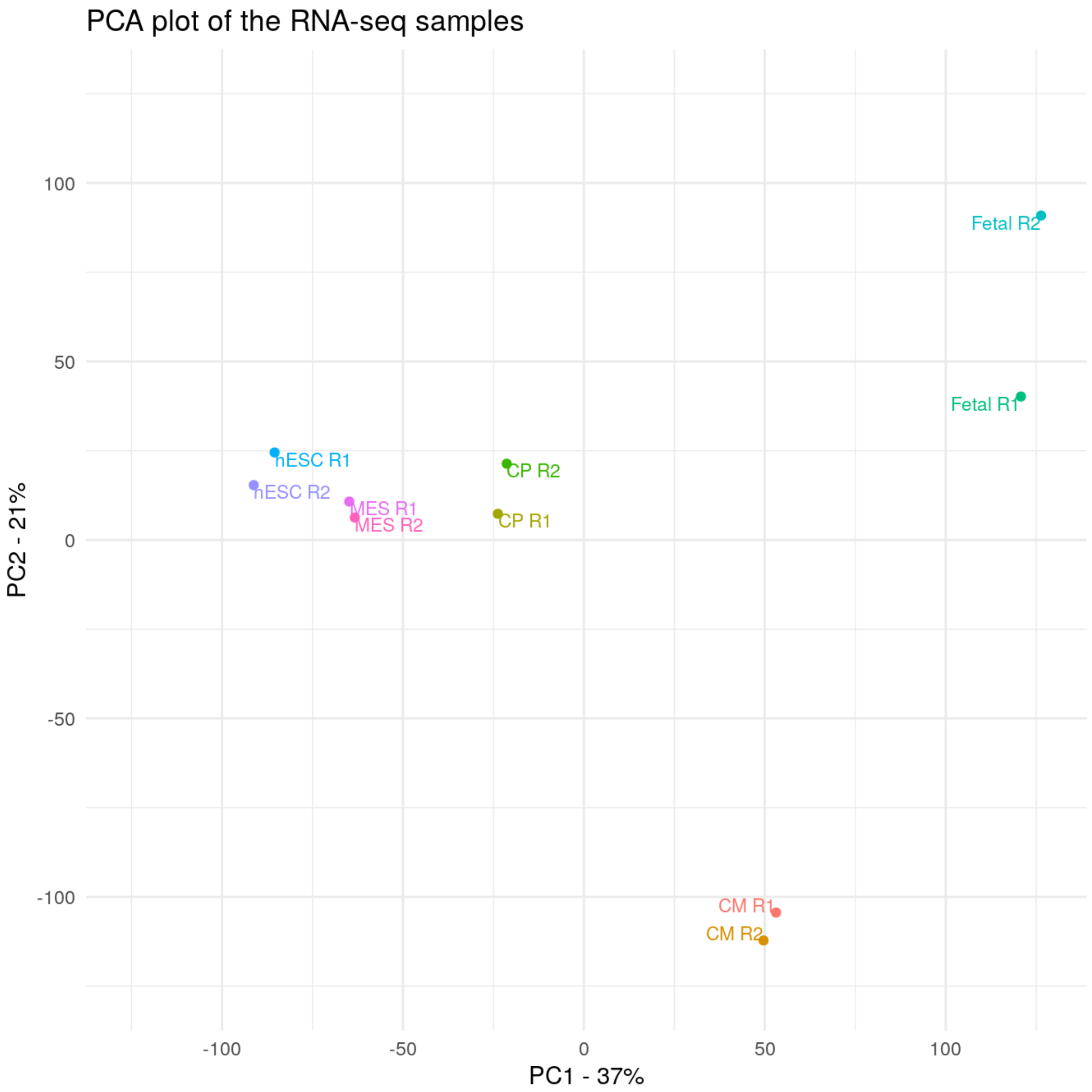
PCA分析过程

```r
pca <- t(as.matrix(genes_expression)) %>%
    prcomp(
        # genes_expression,
        scale = TRUE
    ) # 此处scale参数可不加，使用单独的scale()函数可实现同样效果
var <- data.frame(pca$sdev^2)
var_per <- round(var / sum(var) * 100)
ggbiplot::ggscreeplot(pca) # 碎石图查看方差被解释的部分
```



PCA结果作图

```
data.frame(
    Sample_name = rownames(pca$x),
    X = pca$x[, 1],
    Y = pca$x[, 2]
) %>%
    ggplot(mapping = aes(
        x = X,
        y = Y,
        label = Sample_name,
        color = Sample_name
    )) +
    geom_text(
        vjust = "inward",
        hjust = "inward",
        check_overlap = TRUE,
        size = 3
    ) +
    geom_point() +
    expand_limits(x = c(-125, 125), y = c(-125, 125)) +
    theme_minimal() +
    labs(
        title = "PCA plot of the RNA-seq samples",
        x = paste("PC1 - ", var_per[1, 1], "%", sep = ""),
        y = paste("PC2 - ", var_per[2, 1], "%", sep = "")
    ) +
    theme(
        legend.position = "none"
    )
```



PCA结果分析

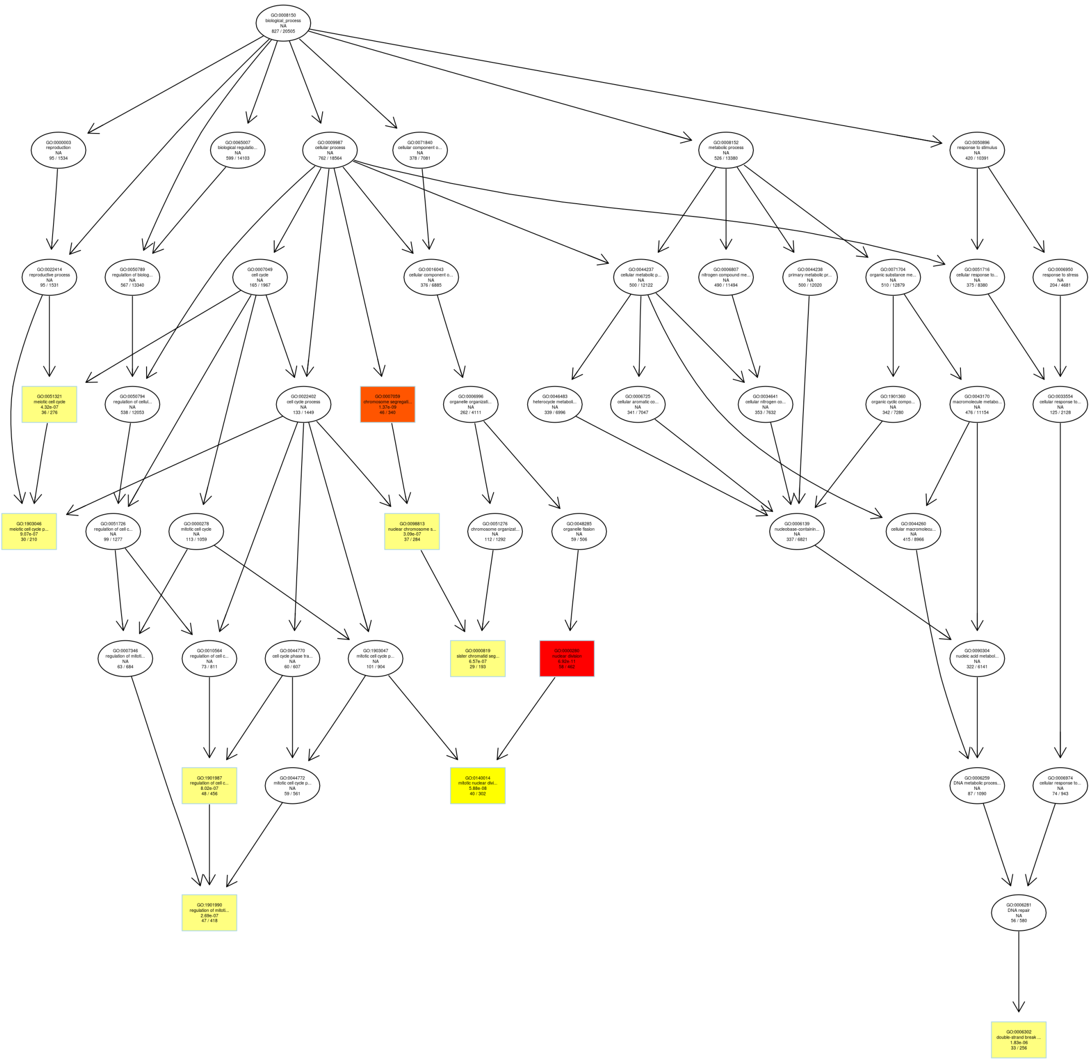
PC1

不难从图上看出，`PC1`主要显示的是细胞类型的分化历程（$hESC \rightarrow MES \rightarrow CP \rightarrow CM \rightsquigarrow Fetal$）

PC2

`PC2`的含义较难从PCA结果图上直观的看出，但可根据以下GO分析结果尽心推测

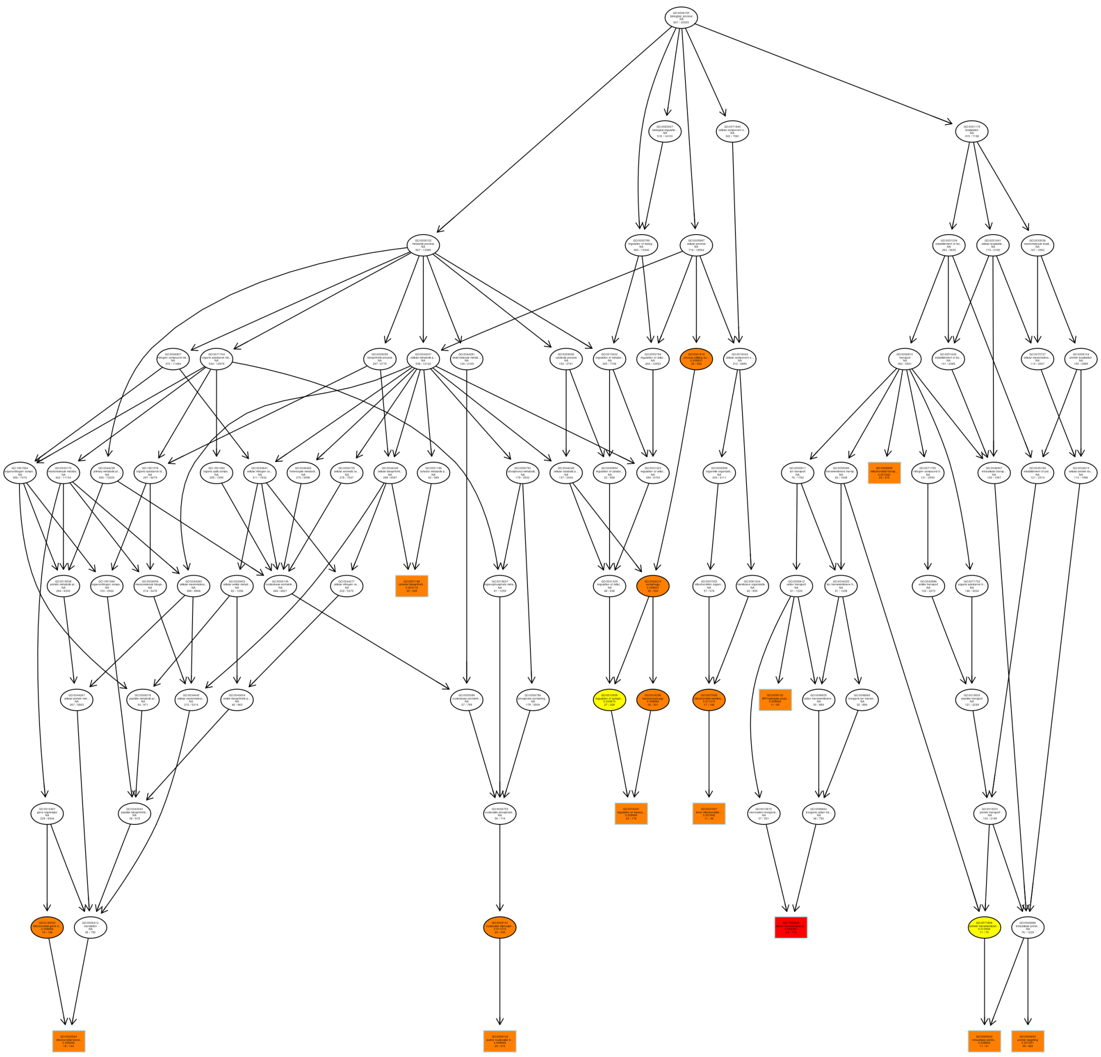
```
library(clusterProfiler)
library(org.Hs.eg.db)
up_bond <- unname(quantile(pca$rotation[, 2], 0.95))
low_bond <- unname(quantile(pca$rotation[, 2], 0.05))
egoP <- enrichGO(
    gene = rownames(pca$rotation[pca$rotation[, 2] > up_bond, ]),
    OrgDb = org.Hs.eg.db,
    keyType = "ENSEMBL",
    ont = "BP",
    pAdjustMethod = "BH",
    pvalueCutoff = 0.01,
    qvalueCutoff = 0.05
)
egoN <- enrichGO(
    gene = rownames(pca$rotation[pca$rotation[, 2] < low_bond, ]),
    OrgDb = org.Hs.eg.db,
    keyType = "ENSEMBL",
    ont = "BP",
    pAdjustMethod = "BH",
    pvalueCutoff = 0.01,
    qvalueCutoff = 0.05
)
plotGOgraph(egoP)
```


GO:0008150 biological_process NA 827 / 20505
GO:0000003 reproduction NA 95 / 1534
GO:0065007 biological regulatio... NA 599 / 14103
GO:0009987 cellular process NA 762 / 18564
GO:0071840 cellular component o... NA 378 / 7081
GO:0008152 metabolic process NA 526 / 13380
GO:0050896 response to stimulus NA 420 / 10391
GO:0022414 reproductive process NA 95 / 1531
GO:0050789 regulation of biolog... NA 567 / 13340
GO:0007049 cell cycle NA 165 / 1967
GO:0016043 cellular component o... NA 376 / 6885
GO:0044237 cellular metabolic p... NA 500 / 12122
GO:0006807 nitrogen compound me... NA 490 / 11494
GO:0044238 primary metabolic pr... NA 500 / 12020
GO:0071704 organic substance me... NA 510 / 12879
GO:0051716 cellular response to... NA 375 / 8380
GO:0006950 response to stress NA 204 / 4681
GO:0051321 meiotic cell cycle 4.32e-07 36 / 276
GO:0050794 regulation of cellul... NA 538 / 12053
GO:0022402 cell cycle process NA 133 / 1449
GO:0007059 chromosome segregati... 1.37e-09 46 / 340
GO:0006996 organelle organizati... NA 262 / 4111
GO:0046483 heterocycle metaboli... NA 339 / 6996
GO:0006725 cellular aromatic co... NA 341 / 7047
GO:0034641 cellular nitrogen co... NA 353 / 7632
GO:1901360 organic cyclic compo... NA 342 / 7280
GO:0043170 macromolecule metabo... NA 476 / 11154
GO:0033554 cellular response to... NA 125 / 2128
GO:1903046 meiotic cell cycle p... 9.07e-07 30 / 210
GO:0051726 regulation of cell c... NA 99 / 1277
GO:0000278 mitotic cell cycle NA 113 / 1059
GO:0098813 nuclear chromosome s... 3.09e-07 37 / 284
GO:0051276 chromosome organizat... NA 112 / 1292
GO:0048285 organelle fission NA 59 / 506
GO:0006139 nucleobase-containin... NA 337 / 6821
GO:0044260 cellular macromolecu... NA 415 / 8966
GO:0007346 regulation of mitoti... NA 63 / 684
GO:0010564 regulation of cell c... NA 73 / 811
GO:0044770 cell cycle phase tra... NA 60 / 607
GO:1903047 mitotic cell cycle p... NA 101 / 904
GO:0000819 sister chromatid seg... 6.57e-07 29 / 193
GO:0000280 nuclear division 6.92e-11 58 / 462
GO:0090304 nucleic acid metabol... NA 322 / 6141
GO:1901987 regulation of cell c... 8.02e-07 48 / 456
GO:0044772 mitotic cell cycle p... NA 59 / 561
GO:0140014 mitotic nuclear divi... 5.88e-08 40 / 302
GO:0006259 DNA metabolic proces... NA 87 / 1090
GO:0006974 cellular response to... NA 74 / 943
GO:1901990 regulation of mitoti... 2.69e-07 47 / 418
GO:0006281 DNA repair NA 56 / 580
GO:0006302 double-strand break ... 1.83e-06 33 / 256

```
$dag
A graphNEL graph with directed edges
Number of Nodes = 51
Number of Edges = 81
```

```
$complete.dag
[1] "A graph with 51 nodes."
```

```
plotGOgraph(egoN)
```



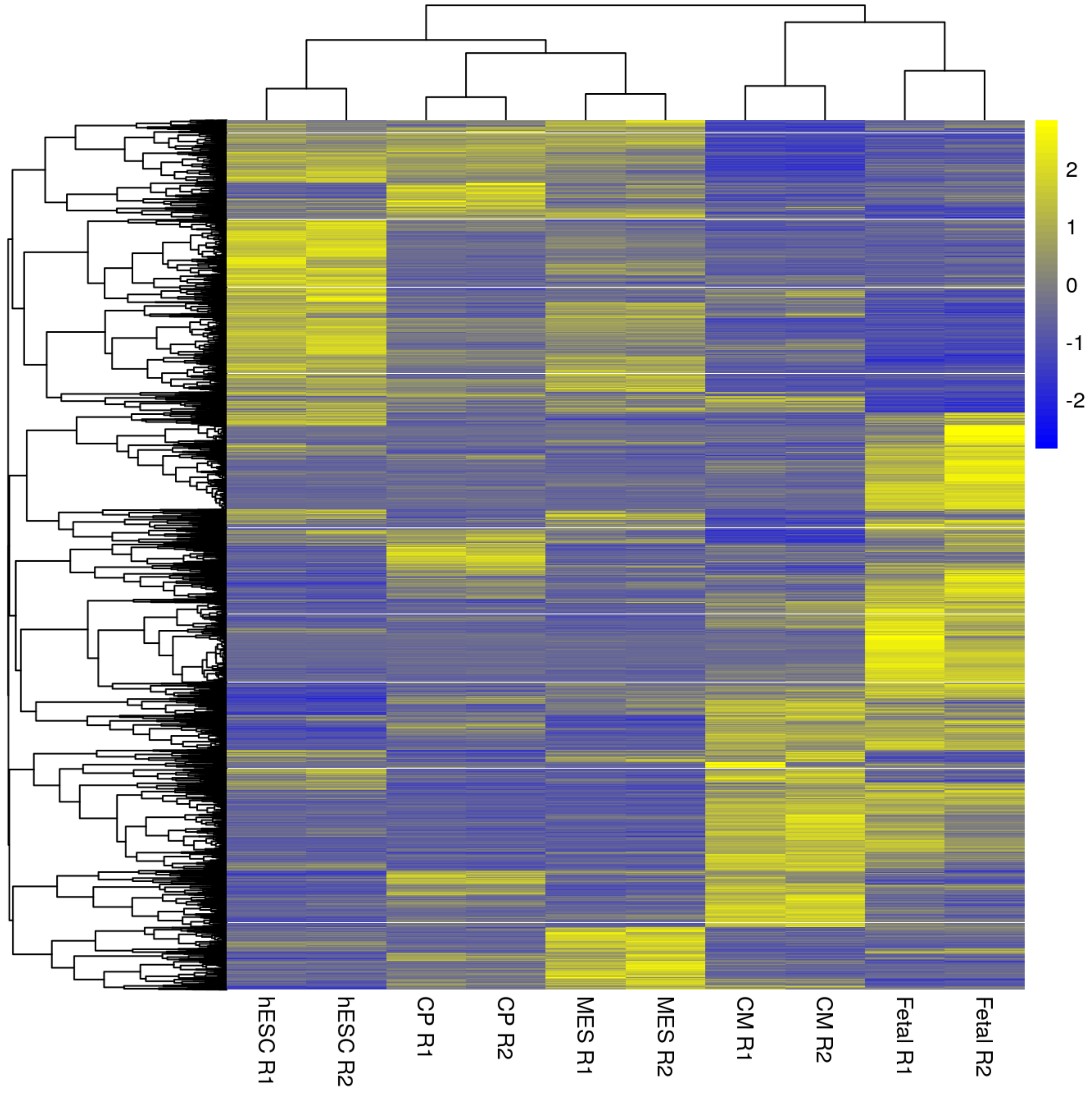
```
$dag
A graphNEL graph with directed edges
Number of Nodes = 94
Number of Edges = 154

$complete.dag
[1] "A graph with 94 nodes."
```

GO结果分析

从图中不难看出，`PC2`中显著上调的基因与染色体分离、细胞核分裂、线粒体活动调控等细胞分裂的过程相关，而对应的显著下调的则是主要集中在质子转运、线粒体转录活动等过程上。因此`PC2`整体反映了细胞的分裂活动的强度，`PC2`值越大，活动强度越高，反之亦然。

差异表达基因热图

```r
library(pheatmap)
signif_diff_gene <- read.table("./DEGs.txt")
signif_diff_gene <- as.vector(signif_diff_gene[[1]])
diff_gene_expression <- genes_expression[rownames(genes_expression) %in%
pheatmap(
    diff_gene_expression,
    clustering_distance_rows = "correlation",
    scale = "row",
    show_rownames = FALSE,
    color = colorRampPalette(c("blue", "yellow"))(50),
    treeheight_row = 100
)
```


hESC R1
hESC R2
CP R1
CP R2
MES R1
MES R2
CM R1
CM R2
Fetal R1
Fetal R2